AFRL-IF-RS-TR-1999-116
Final Technical Report
June 1999

PATTERN RECOGNITION AND IMAGE ANALYSIS EXTENSIONS TO THE IE2000 IPTOOL KIT

University of Vermont

Robert R. Snapp

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-1999-116 has been reviewed and is approved for publication.

APPROVED: Scott F. Adams

SCOTT F. ADAMS
Project Engineer

FOR THE DIRECTOR: John V McNamara

JOHN V. MCNAMARA, Technical Advisor
Information & Intelligence Exploitation Directorate
Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** June 1999	**3. REPORT TYPE AND DATES COVERED** Final Jul 94 - Aug 97

4. TITLE AND SUBTITLE

PATTERN RECOGNITION AND IMAGE ANALYSIS EXTENSIONS TO THE IE2000 IPTOOLKIT

5. FUNDING NUMBERS

C - F30602-94-1-0010
PE - 62702F
PR - 4813
TA - 00
WU - 01

6. AUTHOR(S)

Robert R. Snapp

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Vermont
Department of Computer Science
Burlington VT 13441-4114

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/IFEC
32 Brooks Road
Rome NY 13441-4114

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

AFRL-IF-RS-TR-1999-116

11. SUPPLEMENTARY NOTES

Air Force Research Laboratory Project Engineer: Scott F. Adams/IFEC/(315) 330-1430

12a. DISTRIBUTION AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This research project produced results of both fundamental and practical benefit, including a quantitative description of the finite sample accuracy of the k nearest neighbor classifier for a large family of smooth pattern recognition problems, a new theoretical justification for use of a weighted euclidean metric as a similarity function, the development of a procedure for estimating the Bayes risk of practical problems, and the development of a fast approximation of a k nearest neighbor classifier, called the labeled cell classifier. The research resulted in a stand-alone X Windows software application, called pstool, that allows users to interactively construct training sets from multispectral digital images, six refereed conference publications, a 40 page technical report, and a journal publication in the Annals of Statistics. Seven graduate students at the University of Vermont participated in this project: four receiving Masters degrees, and one, a Ph.D in Electrical Engineering.

14. SUBJECT TERMS

Pattern Recognition, Image Processing, K Nearest Neighbor Classifier, Bayes Classifier

15. NUMBER OF PAGES

28

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Contents

1 Introduction

This report summarizes the research results performed from July 1, 1994 through August 31, 1997 under the project entitled "Pattern Recognition and Image Analysis Extensions to the IE2000 IPToolkit," Grant No. F30602-94-1-0010, funded by the United States Air Force. The research effort was performed at the University of Vermont by Prof. Robert R. Snapp and seven graduate students under his supervision. The work resulted in one journal publication in the *Annals of Statistics*, six conference papers, a 40 page technical report, an X Windows software package, and the development of a novel algorithm that efficiently approximates a k-nearest neighbor classifier in low dimensional feature spaces. The quality of this work is reflected by the high reviewing standards of the journals and conferences used to communicate these results, and the award of a competitive research grant from the U. S. Army Research Office to continue this work[1], and invitations to present talks describing some of these results at colloquia at Cambridge University (Cambridge, England, October 1994), Concordia University (Montreal, Canada, May 1995), Rensselaer Polytechnic Institute (Troy, NY, April 1995), SUNY Buffalo (October 1995), and Siemens Corporate Research, (Princeton, NJ, October 1995). These research results are described in greater detail in the following sections, and complete copies of all publications are contained in Section 6.

Section 2 describes the theoretical results derived, including a finite-sample analysis of the k-nearest neighbor classifier under different metrics, analytic support for the asymptotic optimality of a weighted Euclidean metric. Section 3 describes two algorithms inspired by the theory: (*i*) a strategy for estimating the Bayes risk of a practical pattern classification from a set of classified patterns; and (*ii*) an efficient implementation of the k-nearest neighbor classifier, called the labeled cell classifier. Section 4 describes an X Windows program that incorporates a graphical tool for building labeled and unlabeled reference sets from multispectral images interactively, and includes an efficient implementation of the k-nearest neighbor classifier for classifying other pixels in the same, or related images. Section 5 identifies the students who participated in this research, and were supported on this grant. Section 6 describes the publications that grew out of this project. Section 7 summarizes

[1] "Finite Sample Analyses of Nearest Neighbor Algorithms," U. S. Army Research Office, DAAG55-98-1-0022

and discusses the practical relevance of our work to image exploitation.

2 Theoretical Results

The most significant results of the research were two theoretical discoveries related to a finite-sample analysis of the k-nearest neighbor classifier[12], one of the most popular pattern recognition algorithms in use today. In this context, we assume that each pattern is a vector, constructed from a finite number of measurements, or *features* [10]. As a simple example, each pixel in a multispectral image can be represented as a pattern using the intensities of the spectral bands as features. As a more general example, each pixel can be represented using the intensities obtained from an array of image processing filters that are centered about that pixel (e.g., edge, texture, or shape detectors). In this way, salient information about the values of the neighboring image pixels can be incorporated within a pattern. The number of features (e.g., the sum of the number spectral bands and the number of filters) used to represent each pattern is called the dimensionality of the feature space. In order for this scheme to be useful, features should be selected so that patterns originating from distinct states of nature, or *classes*, are more or less distinguishable. As it is rarely possible in practice to analytically describe how patterns of a given class are generated, almost every classification method is based on the information contained in a training set of correctly labeled patterns, or *reference sample*, that is a set of feature vectors, each labeled by its true class.

Given a reference sample (or "training set") of m labeled (i.e., classified) feature vectors the *k nearest neighbor classifier* assigns an input pattern $\mathbf{x}$ to a class by identifying the subset of k feature vectors from the reference sample that are closest to $\mathbf{x}$ using a predefined distance function (or metric). The input pattern is then assigned to the class that appears most frequently within the subset of k nearest neighbors.

Despite its simplicity, this algorithm has been shown theoretically to be as accurate as a Bayes classifier (the most accurate pattern classifier possible) in the limit of an infinite sample size [35]. Fortunately, this limit converges rapidly for many practical problems, which, along with its ease of use, is why it so popular among practitioners. In image exploitation, nearest neighbor methods can be used to compare the accuracy of different feature representations for a given classification problem. Because there is no extensive training phase, k nearest neighbor classifiers can be quickly constructed and put on line, as new application needs arise.

2.1 Finite-Sample Analysis

The research that we performed provides a quantitative description on how this limit is achieved, and enables an improved understanding of this classifier's performance using finite reference samples. This work extends the classic results of Cover and Hart [5, 4], and corrects the recent work of Fukunaga and Hummels [19]. Specifically, in a series of papers [28, 29, 30, 31] we showed that for classification problems that possess a certain degree of regularity, the probability of error of the k nearest neighbor classifier $P_m(\text{error})$ can be accurately estimated from an asymptotic series of the form

$$P_m(\text{error}) = c_0 + \sum_{j=2}^{\infty} c_j m^{-j/n}. \tag{1}$$

Here, $c_0 = P_\infty(\text{error})$ denotes the expression derived by Cover and Hart [5] for the probability of error of the k nearest neighbor classifier in the infinite sample limit; m denotes the number of labeled patterns in the reference sample, and n, the dimensionality of each pattern. As (1) is an asymptotic expansion in the sense of Poincaré [11], it can be truncated at any point, resulting in an error with magnitude of the first neglected term. We also have obtained analytic expressions for the leading expansion coefficients, c_j, in the summation of Eqn. (1) in terms of the probability distributions that define the pattern recognition problem, the value of k, and the metric used.

Eqn. (1) is significant for the following reasons:

- If the probability distributions that describe each pattern class are known, an N-th order truncation of (1) can be used to predict the finite sample accuracy of a k nearest neighbor classifiers, as values of the expansion coefficients $P_\infty(\text{error}), c_2, c_3, \ldots, c_N$ can be evaluated numerically from the expressions published in [31].
- If the probability distributions that describe each pattern class are *not* known, as is the case in nearly every pattern recognition problem of practical interest, then an N-th order truncation of (1) can be used to predict the finite sample accuracy of a k nearest neighbor classifier, as values of the expansion coefficients $P_\infty(\text{error}), c_2, c_3, \ldots, c_N$ can be

estimated statistically using standard resampling methods [27, 33]. In particular, (1) can be used to estimate the practical benefit of acquiring more reference data, and thus is useful for designing k nearest neighbor classifiers. (Section 3.1 describes an extension of this idea for estimating the Bayes risk from a labeled reference sample.)

- Eqn. (1), and the analytic form of the leading coefficients, provide useful fundamental insights about this algorithm. For example, the factor of $m^{-2/n}$ in the second term is an analytic validation of the curse of dimensionality. Similarly an analysis of coefficient c_2, described below, demonstrates that a weighted Euclidean metric is asymptotically optimal for the class of problems considered by this analysis.

2.2 Asymptotic optimality of the Euclidean metric

A pressing issue in the realm of applied pattern recognition is how does one design a pattern classifier for a given problem. In the context of the k nearest neighbor algorithm, one might ask what metric yields the most accurate classifier. This is an open problem, and generally depends upon specifics of the problem. Nevertheless, for the class of sufficiently smooth problems our work demonstrates that for sufficiently large sample size, a weighted Euclidean metric is the optimal global L_p metric [30, 31].

To show this, we considered a broad class of global metrics derived from the standard L_p norm:

$$\|\mathbf{x}\|_p = \begin{cases} \sqrt[p]{|x_1|^p + \cdots + |x_n|^p} & : \text{ if } 1 \le p < \infty, \\ \max\limits_{1 \le i \le n} |x_i| & : \text{ if } p = \infty, \end{cases}$$

and assumed the general global metric

$$d(\mathbf{x}, \mathbf{y}) = \|A(\mathbf{x} - \mathbf{y})\|_p$$

where A is an arbitrary nonsingular n-by-n matrix, and p is chosen from the interval $1 \le p \le \infty$. Under these assumptions, we showed that Eqn. (1) converges uniformly with respect to values of A and p. Since the leading coefficient c_0 does not depend on these values, the optimal asymptotic metric can be found by finding the values of A and p that minimize the next most significant coefficient, namely c_2. Surprisingly, the optimal value of p

equals 2, independent of the specifics of the pattern recognition problem. We also obtained an expression for the optimal weight matrix A in terms of the probability distributions that define the given pattern recognition problem. Numerical simulations were also used to demonstrate the practical significance of these findings [31].

3 Algorithms Developed

Two promising algorithms were developed during the course of this project. The first directly stems from the theoretical analysis described in the previous section for estimating the Bayes risk of a practical pattern recognition problem from real data [27, 33]. The second algorithm, the labeled cell classifier, is a computationally efficient approximation to a k nearest neighbor classifier [24]. Both algorithms are described below.

3.1 Estimating the Bayes risk

Given a pattern classification problem, a *Bayes classifier* is defined to be a pattern classifier that minimizes the probability of error (or in more general terms the statistical risk, as some misclassifications may incur a greater cost than others). Computationally, a Bayes classifier assigns each input pattern to the class that has the maximum posterior probability [10]. The probability of error of such a classifier is called the *Bayes risk*; we shall denote its value by R_B.

In practice the construction of a Bayes classifier generally requires knowledge of the probability distributions that define the given pattern classification problem. Unfortunately, this information is usually not available for problems of practical interest. Nevertheless, our research demonstrates that accurate estimates of the Bayes risk can be obtained from a sufficiently large reference sample of labeled feature vectors. Estimates of the Bayes risk can facilitate the design of better classifiers. For example, since the value of R_B depends upon the set of features chosen to represent each pattern, one might compare estimates of the Bayes risk for a number of different feature sets, and then select the representation that yields the smallest value. Several previous efforts have utilized the k nearest neighbor classifier towards this end [6, 20, 21, 23]. Our work reported in references [27, 33] takes advantage of Eqn. (1), the most detailed parametric model available of the accuracy of this pattern classifier as a function of the reference sample size m. Using a large pool of classified data, one can construct a sequence of k nearest neighbor classifiers of varying sample sizes. Using standard least squares methods, one can obtain estimates for the unknown coefficients $c_0, c_2, \ldots, c_N$. Of greatest utility is the estimated value of c_0 which can be used to place upper and lower bounds on the value of the Bayes risk R_B: To enable an estimate of R_B with precision ϵ, choose $k > 2/\epsilon^2$, and estimate c_0 by the above method. After inverting an inequality derived by Devroye [7] one obtains

$$c_0 - \epsilon \leq R_B \leq c_0.$$

The practical utility of this method was demonstrated by two experiments, one using synthetic data, where the true Bayes risk was known, and the other using a pattern classification problem using imagery obtained from the IE2000 group at Rome Laboratory (See Fig. 1).

3.2 Labeled cell classifier

Although accurate and easy to adapt to new classification problems, the time that a k nearest neighbor classifier needs to classify an input pattern increases with the size of the reference sample. Over the years, different techniques have been developed to simplify the search for the k nearest neighbors in a reference sample. (That the *Sixth DIMACS Implementation Challenge* for 1998 centers on this task, demonstrates that this remains a ripe problem [9].) Most approaches fall within two broad categories: (*i*) exact implementations that restrict the search for nearest neighbors, by organizing the data in a hierarchal manner [13, 15, 17, 14], or (*ii*) approximate implementations that edit the reference sample [1, 16, 36]. In our study [24], we contructed a hybrid approach.

3.2.1 *k*-d trees

The labeled cell algorithm, described in Section 3.2.2, is based on the implementation of Friedman, Bentley, and Finkel[14] that organizes the reference sample $\mathcal{X}_m$ into an n-dimensional binary tree, such that the root node represents the entire feature space, and each node in the tree represents an isothetic cell that contains a subset of $\mathcal{X}_m$. The two descendants of each nonterminal node divide the parent cell along one coordinate, called the *key*,

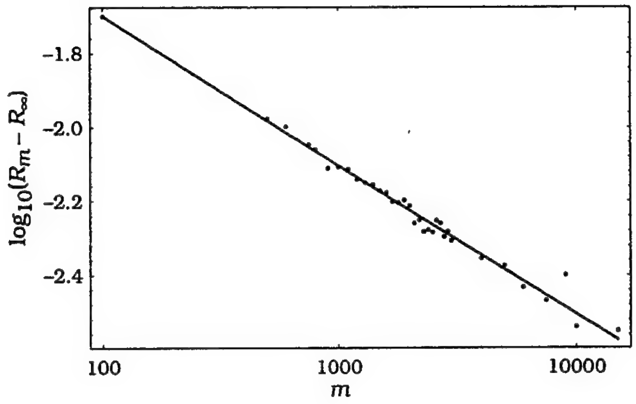

Figure 1: A fourth-order ($N = 4$) fit of Eqn. (1) to 33 empirical estimates of $\hat{R}_{m_i}$ for a pixel classification problem obtained from a multispectral satellite image. Patterns were constructed using five spectral components of each image pixel. Using $R_\infty = 0.0758$, the fourth-order fit, $R_m = 0.0758 + 0.124m^{-2/5} + 0.0133m^{-4/5}$, is plotted as a solid curve on a log-log scale to reveal the significance of the $j = 2$ term.

such that the number of reference patterns in each child cell differs at most by one. The key may be the coordinate of greatest variation of the reference vectors in the parent cell, and the threshold may be the median of their projections along the chosen coordinate. Pairs of descendants are added recursively until the number of vectors in a cell does not exceed a bucket size b. Note that the nodes at a constant depth represent a partition of the feature space, as do the leaf nodes. Fig. 2 displays a k-d tree constructed from a reference sample of 16 points in $\mathbb{R}^2$, with $b = 2$.

After the tree is completed, the k feature vectors in the tree that are nearest to a given input pattern $\mathbf{x}$ can be identified. A priority queue is used to maintain the k feature vectors encountered so far that are closest to $\mathbf{x}$. Beginning with the root, nodes in the tree are examined recursively until it is certain that the k nearest neighbors have been found. If the current node

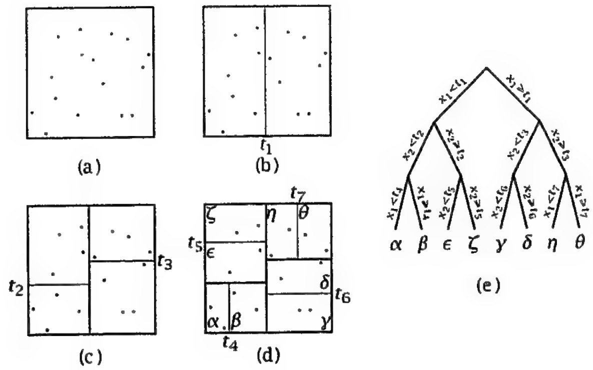

Figure 2: (*a*) A k-d tree of depth three is constructed from this set of sixteen feature vectors in $\mathbb{R}^2$ that forms the root of the tree. (*b*) The set is bisected into left and right portions, forming the two descendent nodes of the root, as the largest variation appears along the horizontal coordinate. (*c*) Each resulting subset is further divided into two equal partitions along the vertical coordinate, forming the four nodes at depth two in the tree. (*d*) Each resulting subset is then divided along the coordinate of greatest variation. Each resulting cell, labeled with a greek letter, contains two feature vectors, and forms a leaf node of the k-d tree (*e*).

is a leaf node, then the priority queue is updated after examining its b or fewer feature vectors. Otherwise the key i and threshold value t of the node are examined, and the recursive procedure is applied first to the descendant that falls on the same side of t as x_i, and then to its sibling. For efficiency, nodes are only examined if their cell boundaries are closer to $\mathbf{x}$ than the k-th nearest neighbor found so far (the bounds-overlap-ball test); and the search is stopped as soon as the k-th nearest neighbor is closer to $\mathbf{x}$ than the boundaries of every unexamined cell (the ball-within-bounds test).

3.2.2 Labeling the cells

The labeled cell algorithm is designed to reduce the number of feature vectors examined during each classification. As in the previous implementation, the reference sample is organized into a multidimensional binary search tree using the coordinates of the feature vectors as keys. An integer $k' \geq k$ and a fraction $\alpha > 0$ are selected. A central test vector from each leaf cell is then classified with an exact k'-nearest-neighbor classifier (e.g., the previous implementation). This test vector could be the centroid of the leaf cell (assuming it is compact), or the sample mean of its reference vectors. If the number of k'-nearest-neighbors that belong to the most frequent class exceeds $\lfloor \alpha k' \rfloor$, then the leaf cell is given the label of that class.[2] (Otherwise, it remains unlabeled.) Nonterminal nodes are examined recursively: if two siblings share a common class label, then their parent is assigned the same label.

Input patterns are classified by the k-d tree algorithm, with one important exception: if an input pattern belongs to a cell that is labeled, then it is immediately assigned to the indicated class. Thus no reference vectors are examined if an input falls within a labeled cell. For different values of α, k', and k, the labeled cell algorithm implements a variety of classifiers: $\alpha = 1$ yields an exact k-nearest neighbor classifier, and $\alpha \leq 1/C$, a pure cell classifier.

Since the classes assigned to patterns that fall within the labeled cells may differ occasionally from the results of the k-nearest neighbor algorithm. Thus, like Hart's condensed nearest neighbor rule [16], the labeled cell classifier only approximates the classic algorithm. However, computer experiments suggest that if a classification needs to be performed in a fixed amount of time, then the new algorithm may attain greater accuracy than other implementations of the k-nearest-neighbor classifier, as the computation saved in the labeled cells allows this new algorithm to process a larger reference sample.

3.2.3 Experimental results

Two problems illustrate the differences in performance and accuracy between labeled cell and exact k-d tree implementations of the k-nearest neighbor classifier. The first, assumes two equally probable, normally distributed classes in $\mathbb{R}^3$. Thus the class-conditional probability densities are

$$f_\ell(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{3/2}} e^{-((x_1+(-1)^\ell\mu)^2+x_2^2+x_3^2)/2\sigma^2},$$

for $\ell \in \{1,2\}$. The classification accuracy (i.e., the expected probability of error), and the expected number of operations per classification are empiri-

[2] For simplicity, we assume a zero-one loss matrix, so cells are labeled if their local estimate of the conditional risk is less than $1 - \alpha$. It is straightforward to generalize the algorithm to a asymmetric, multiclass, risk function.

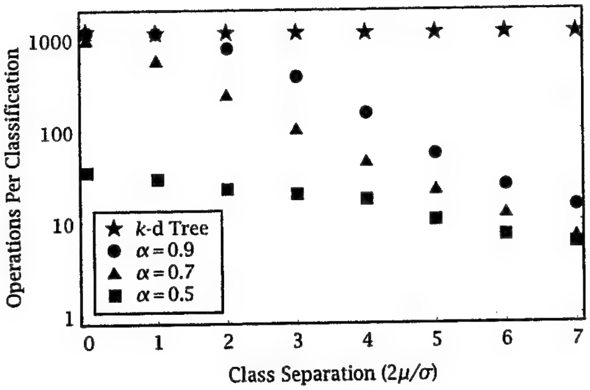

Figure 3: A semilogarithmic plot obtained from a classification problem with two normally distributed classes in $\mathbb{R}^3$. The circular, triangular, and square markers describe the average performance of hundreds of labeled cell classifiers with α equal to 0.9, 0.7, and 0.5 (a pure cell classifier) respectively. In all cases $k = k' = 11$. The five-pointed stars describe the performance of an k-d tree implementation of an 11-nearest-neighbor classifier. Vertical error bars all lie within each marker.

cally estimated from a sequence of independent trials. For each trial a random reference sample of $m = 10,000$ patterns is used to classify several thousand independent input vectors. The number of operations is estimated heuristically: each comparison and addition count as one operation, and each multiplication as two. (Qualitatively similar results are obtained with a variety of weighting factors.) Results for $k = 11$, a Euclidean metric, and eight values of $2\mu/\sigma$ are displayed in Fig. 3. In this example, the greatest absolute deviation in accuracy between two implementations occurs at $2\mu/\sigma = 6$ and $\alpha = 0.5$, where the labeled cell classifier misclassifies 0.15% of the independent test patterns, and the k-d tree implementation misclassifies 0.14%. Note in particular, how the recursive labeling scheme accelerates the performance as the class separation is increased, with little degradation in accuracy.

The second problem, uses data extracted from a seven-band digital image. We let each pixel define an independent pattern. The first band is quantized about the median to obtain a binary class label. A six-dimensional feature vector is formed with the remaining spectral bands. Reference and test patterns are selected independently from the image. Fig. 4 displays the trade-off between the classification accuracy and the computational cost for four different reference sample sizes as well as four different values of α. These

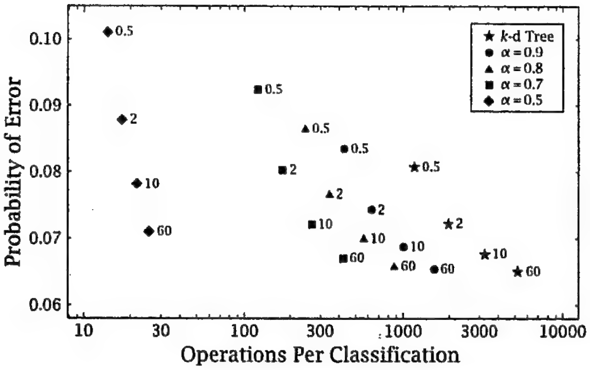

Figure 4: Results of the second experiment in which six-dimensional pixels, belonging to two different classes were classified by three different labeled cell classifiers ($k' = k = 7$), and a k-d tree implementation of a 7-nearest-neighbor classifier. The reference sample size appears to the right of each marker in thousands. The horizontal axis is logarithmic.

results suggest that the recursive labeling scheme accelerates classification with only a small reduction in accuracy. Note that by increasing the size of the reference sample, it is possible to obtain a labeled cell classifier that is both significantly faster and more accurate than a k-d tree classifier. Thus the new algorithm may be useful for real-time applications that provide an abundant supply of classified data. The estimates, redisplayed in Fig. 5, validate that the average classification time of labeled cell classifiers is also $O(\log m)$, but with smaller constants of proportionality β. Preliminary comparative experiments suggest that the labeled cell classifier is competitive with other approximations of the k nearest neighbor algorithm. Moreover, recursive labeling can be combined with early truncation (Arya and Mount [1]) to yield even faster implementations.

These simulations suggest that the labeled cell classifier is most useful for problems that provide an abundant supply of classified patterns, are described by smooth probability distributions, and have a small Bayes risk (*e.g.*, pixel classification of satellite images).

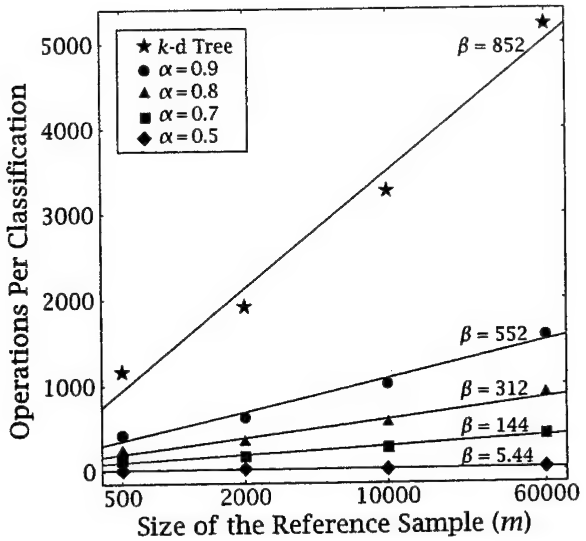

Figure 5: Empirical estimates of the average number of operations required for each classification as a function of the sample size m for the second experiment. The linear graphs represent least-square fits of the form $\beta \log_{10} m + \gamma$. (Note that the horizontal axis is logarithmic.)

4 Software Production

The most successful software to emerge from this project is a stand-alone X Windows application called *pstool* that enables a user to interactively construct a labeled reference sample from a multispectral digital image (in either LANDSAT-TM or TIFF format) and classify other regions of the image using a k-nearest neighbor classifier. This application was brought to Rome Labs for a demo in August 1995, and a revised version was placed on an FTP server in the Spring of 1996. The program was also used by an image processing group at Rensselaer Polytechnic Institute in Troy, New York. The program was written in C in a modular fashion, using updated IPToolkit modules.

The program proved to be useful for our following experimental research in Bayes risk estimation, and in designing faster implementations of k nearest neighbor classifiers.

Several graduate students also contributed software to the project, including a C++ class library of three different neural network training algorithms, an implementation of the time-difference reinforcement learning algorithm, and algorithms for detecting roads in digital images.

5 Students Supervised

This grant help further the education and professional training of seven graduate students at the University of Vermont, four of whom received Master of Science degrees, and one received a Ph.D. Students benefited from research assistantships awarded during the summer and for Mr. Yong Feng, during the academic year. Their names are listed below, along with the degrees they received.

- Mr. Tong Xu, M.S. in Electrical Engineering, 1995.
- Dr. Alessandro Palau, Ph.D. in Electrical Engineering, 1997.
- Mr. Xianguan Li, M.S. in Electrical Engineering, 1997.
- Mr. Yong Feng, M.S. in Computer Science, 1997.
- Mr. Chaoyu Jin, M.S. in Electrical Engineering, 1997.
- Mr. Qing Ye, graduate student in Computer Science
- Mr. Shawn Ma, graduate student in Computer Science

6 List of Publications

The most significant results of this research project appear in seven papers: six were accepted by peer reviewed conferences, and one, by the *Annals of Statistics*, the flagship and stringently reviewed journal of the Institute of Mathematical Statistics. Copies of these papers appear in the Appendix of this report. An eighth paper, with Alessandro M. Palau on the labeled cell classifier, is in progress.

6.1 Accepted Papers

(a) R. R. Snapp and S. S. Venkatesh, "Asymptotic predictions of the finite-sample risk of the k-nearest-neighbor classifier," *Proceedings of the*

12th International Conference on Pattern Recognition, vol. 2, IEEE Computer Society Press: Los Alamitos, CA, pp. 1-7, 1994.

(b) R. R. Snapp and S. S. Venkatesh, "The finite-sample risk of the k-nearest-neighbor classifier under the L_p metric," *Proceedings of the 1994 IEEE-IMS Workshop on Information and Statistics*, (Alexandria, VA), IEEE Service Center, Piscataway, NJ, October 1994, p. 98.

(c) R. R. Snapp, "Predicting the accuracy of Bayes classifiers," in K. M. Hanson and R. N. Silver, ed., *Maximum Entropy and Bayesian Methods: Sante Fe, New Mexico, U.S.A., 1995* Kluwer Academic Publishers, Dordrecht, Netherlands, 1996, pp. 295–302.

(d) R. R. Snapp and S. S. Venkatesh, "k Nearest Neighbors in Search of a Metric" *Proceedings of the 1995 IEEE International Symposium on Information Theory*, (Whistler, Canada), September 1995.

(e) R. R. Snapp and T. Xu, "Estimating the Bayes Risk from Sample Data," in D. S. Touretzky, M. C. Moser, and M. E. Hasselmo, ed., *Advances in Neural Information Processing Systems*, vol. 8, Cambridge, MA: MIT Press, 1996, pp. 232–238.

(f) A. M. Palau and R. R. Snapp, "The labeled cell classifier: a fast approximation to k nearest neighbors," in A. K. Jain, S. Venkatesh, and B. C. Lovell, ed., *Proceedings of the 14th International Conference on Pattern Recognition*, vol. 1, IEEE Computer Society Press: Los Alamitos, CA, 1998, pp. 823–827.

(g) R. R. Snapp and S. S. Venkatesh, "Asymptotic expansions of the k-nearest neighbor risk," *Annals of Statistics*, vol. 26, no. 3, pp. 850-878,1998.

(h) R. R. Snapp and S. S. Venkatesh, "Asymptotic derivation of the finite-sample risk of the k nearest neighbor classifier," Technical Report UVM-CS-1998-0101, Department of Computer Science, University of Vermont, 1998 (40 pages)

6.2 Manuscripts

(i) A. M. Palau and R. R. Snapp, "The labeled cell classifier," (in progress, to be submitted to *IEEE Trans. Pattern Anal. and Mach. Intell.*).

7 Summary and Practical Consequences

For the image analyst, and indeed any practitioner of pattern recognition, the art of pattern recognition has been, and continues to be, an empirical science. Algorithms are evaluated on their efficiency and accuracy when applied to the problems of interest. Because of the diverse structure of practical classification problems, even in the context of image analysis, it seems unlikely that their exists a unique optimal classification algorithm [8]. However, for many applications the k nearest neighbor algorithm serves as a nearly optimal practical pattern classifier. For example, it is the most popular classification algorithm in handwritten document analysis [25, 34], and a competitive benchmark in general [3].

The results of this study should help practitioners in every field, including image analysis, make better use of the k nearest neighbor classifier. First of all, the asymptotic analysis described by Eqn. (1) (see also [31, 32]), provides a parametric model of the accuracy of this classifier in terms of the reference sample size. In two conference articles [27, 33] (see Figure 1) we demonstrated that this model is valid in the context of pixel classification in multispectral images. Thus, the practitioner can use Eqn. (1), with the least squares technique described in [27, 33], to predict the accuracy of the k nearest neighbor classifier for a range of sample sizes. This information should help answer the question, "How large a reference sample should I use to obtain a pattern classifier that is accurate to within x% of the asymptotic limit?"

Our study also demonstrates analytically how the accuracy of the k nearest neighbor classifier can be enhanced by the selection of an appropriate metric, or distance function. We have shown that for a large class of problems, the choice of a weighted Euclidean metric is the optimal global L_p metric. Future research based upon on this work, may yield methods for discovering the optimal *local* metric directly from the reference data. This will allow practitioners to design more accurate nonparametric pattern classifiers for practical problems. The benefits of this line of research should be most pronounced for classification problems in high dimensional feature spaces [22], such as those encountered in the contexts of multispectral and hyperspectral image analysis.

The main intent of [27, 33] was to demonstrated how this model can be inverted to obtain a estimates of the accuracy of the Bayes classifier for practical pattern classification problems. This knowledge allows the practitioner to compare the intrinsic accuracy of competing representations of a given classification problem. The question "Which spectral bands and image processing filters should I use to represent patterns for identifying objects of class x in environment y?" is an instance of the problem of feature selection, which remains the most important (and perhaps the most difficult) unsolved problem in the field of pattern recognition.

The labeled cell classifier, described above, provides an accurate approximation to the k nearest neighbor classifier in applications where the classification time is critical. As such situations seem likely to occur in defense applications, this algorithm should be of interest to the Air Force.

Finally, the software extensions that we have developed, have provided a useful bridge between our theoretical and empirical investigations, allowing us to efficiently construct labeled reference samples pixel based patterns from multispectral and TIFF images.

8 Bibliography

References

[1] S. Arya and D. M. Mount, "Approximate nearest neighbor queries in fixed dimensions," in *Proceedings of the 4th ACM-SIAM Symposium on Discrete Algorithms*, 1993, pp. 271–280.

[2] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, 1975, pp. 509–517.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth & Brooks: Pacific Grove, CA, 1984.

[4] T. M. Cover, "Rates of convergence of nearest neighbor decision procedures," *Proc. First Annual Hawaii Conf. on Systems Theory*, pp. 413–415, 1968.

[5] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 21-27, 1967.

[6] P. A. Devijver, "A multiclass, $k-NN$ approach to Bayes risk estimation," *Pattern Recognition Letters*, vol. 3, 1985, pp. 1-6.

[7] L. Devroye, "On the asymptotic probability of error in nonparametric discrimination," *Annals of Statistics*, vol. 9, 1981, pp. 1320-1327.

[8] L. Devroye, L. Györfi, G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York, 1996.

[9] Information about the *Sixth DIMACS Implementation Challenge* (1998) can be obtained from the URL `http://dimacs.rutgers.edu/Challenges/Sixth/`.

[10] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York, New York: John Wiley & Sons, 1973.

[11] A. Erdélyi, *Asymptotic Expansions.* New York, New York: Dover, 1956.

[12] E. Fix and J. L. Hodges, Jr., "Discriminatory Analysis — Nonparametric Discrimination: Consistency Properties," *Project 21-49-004, Report No. 4*, USAF School of Aviation Medicine, Randolf Field, TX, 1951, pp. 261-279.

[13] J. H. Friedman, F. Baskett, and L. J. Shustek, "An Algorithm for Finding Nearest Neighbors," *IEEE Trans. Comput.*, vol. C-24, 1975, pp. 1000-1006.

[14] J. H. Friedman, J. L. Bentley, and R. A. Finkel, "An algorithm for finding best matches in logarithmic expected time," *ACM Transactions on Mathematical Software*, vol. 3, 1977, pp. 209-226.

[15] K. Fukunaga and P. M. Narendra, "A Branch and Bound Algorithm for Computing k-Nearest Neighbors," *IEEE Trans. Comput.*, vol. C-24, 1975, pp. 750-753.

[16] P. E. Hart, "The condensed nearest neighbor rule," *IEEE Trans. Inform. Theory*, vol. IT-1, 1968, pp. 515-516.

[17] B. S. Kim and S. B. Park, "A fast k nearest neighbor finding algorithm based on the ordered partition," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. PAMI-8, 1986, pp. 761-766.

[18] K. Fukunaga and T. E. Flick, "An optimal global nearest neighbor metric," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. PAMI-6, pp. 314-318, 1984.

[19] K. Fukunaga and D. M. Hummels, "Bias of nearest neighbor estimates," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. PAMI-9, pp. 103-112, 1987.

[20] K. Fukunaga and D. Hummels, "Bayes error estimation using Parzen and k-NN procedures," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 9, 1987, pp. 634-643.

[21] J. M. Garnett, III and S. S. Yau, "Nonparametric estimation of the Bayes error of feature extractors using ordered nearest neighbor sets," *IEEE Transactions on Computers*, vol. 26, 1977, pp. 46-54.

[22] T. Hastie and R. Tibshirani, "Discriminant Adaptive Nearest Neighbor Classification," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 18, 1996, pp. 607-615.

[23] G. Loizou and S. J. Maybank, "The nearest neighbor and the Bayes error rate," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 9, 1987, pp. 254-262.

[24] A. M. Palau and R. R. Snapp, "The labeled cell classifier: a fast approximation to k nearest neighbors," in A. K. Jain, S. Venkatesh, and B. C. Lovell, ed., *Proceedings of the 14th International Conference on Pattern Recognition*, vol. 1, IEEE Computer Society Press: Los Alamitos, CA, 1998, pp. 823-827.

[25] S. J. Smith, M. O. Bourgoin, K. Sims, H. L. Voorhees, "Handwritten character classification using nearest neighbor in large databases," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. PAMI-16, pp. 915-919, 1994.

[26] D. Psaltis, R. R. Snapp, and S. S. Venkatesh, "On the finite sample performance of the nearest neighbor classifier," *IEEE Transactions on Information Theory* 40 (1994), pp. 820-837.

[27] R. R. Snapp, "Predicting the accuracy of Bayes classifiers," in K. M. Hanson and R. N. Silver, ed., *Maximum Entropy and Bayesian Methods: Sante Fe, New Mexico, U.S.A., 1995* Kluwer Academic Publishers, Dordrecht, Netherlands, 1996, pp. 295–302.

[28] R. R. Snapp and S. S. Venkatesh, "The finite-sample risk of the k-nearest-neighbor classifier under the L_p metric," *Proceedings of the 1994 IEEE-IMS Workshop on Information and Statistics*, (Alexandria, VA), IEEE Service Center, Piscataway, NJ, October 1994, p. 98.

[29] R. R. Snapp and S. S. Venkatesh, "Asymptotic predictions of the finite-sample risk of the k-nearest-neighbor classifier," *Proceedings of the 12th International Conference on Pattern Recognition*, vol. 2, IEEE Computer Society Press: Los Alamitos, CA, pp. 1–7, 1994.

[30] R. R. Snapp and S. S. Venkatesh, "k Nearest Neighbors in Search of a Metric" *Proceedings of the 1995 IEEE International Symposium on Information Theory*, (Whistler, Canada), September 1995.

[31] R. R. Snapp and S. S. Venkatesh, "Asymptotic expansions of the k-nearest neighbor risk," *Annals of Statistics*, 1998.

[32] R. R. Snapp and S. S. Venkatesh, "Asymptotic derivation of the finite-sample risk of the k nearest neighbor classifier," Technical Report UVM-CS-1998-0101, Department of Computer Science, University of Vermont, 1998 (40 pages).

[33] R. R. Snapp and T. Xu, "Estimating the Bayes Risk from Sample Data," in D. S. Touretzky, M. C. Moser, and M. E. Hasselmo, ed., *Advances in Neural Information Processing Systems*, vol. 8, Cambridge, MA: MIT Press, 1996, pp. 232–238.

[34] Sargur N. Srihari, Center for Excellence in Document Analysis, SUNY at Buffalo, Personal Communication, 1996.

[35] C. J. Stone, "Consistent nonparametric regression," *Annals of Statistics*, vol. 5, 1977, pp. 595–645.

[36] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Trans. Systems, Man, and Cybernetics*, vol. SMC-2, 1972, pp. 408–421.

MISSION
OF
AFRL/INFORMATION DIRECTORATE (IF)

The advancement and application of information systems science and technology for aerospace command and control and its transition to air, space, and ground systems to meet customer needs in the areas of Global Awareness, Dynamic Planning and Execution, and Global Information Exchange is the focus of this AFRL organization. The directorate's areas of investigation include a broad spectrum of information and fusion, communication, collaborative environment and modeling and simulation, defensive information warfare, and intelligent information systems technologies.